Klasyfikacja danych opisanych za pomocą szeregów czasowych

Promotor: dr inż. Jerzy Błaszczyński

Realizacja: Wojciech Mioduszewski

Cele

- **Cel**
  - Opracowanie i implementacja różnych podejść do klasyfikacji danych czasowych.
- **Zadania:**
  - Zapoznać się z literaturą tematu.
  - Opracować wybrane podejścia do klasyfikacji danych czasowych.
  - Zaimplementować i udokumentować zaproponowane rozwiązania.
  - Przeprowadzić eksperyment obliczeniowy

Czym jest szereg czasowy

Szereg czasowy *to realizacja procesu stochastycznego, którego dziedziną jest czas*

- Ciąg informacji uporządkowanych w czasie
- Krok może być regularny, bądź nie – (szeregi czasowe rozmyte)

Sposoby analizy przebiegów czasowych

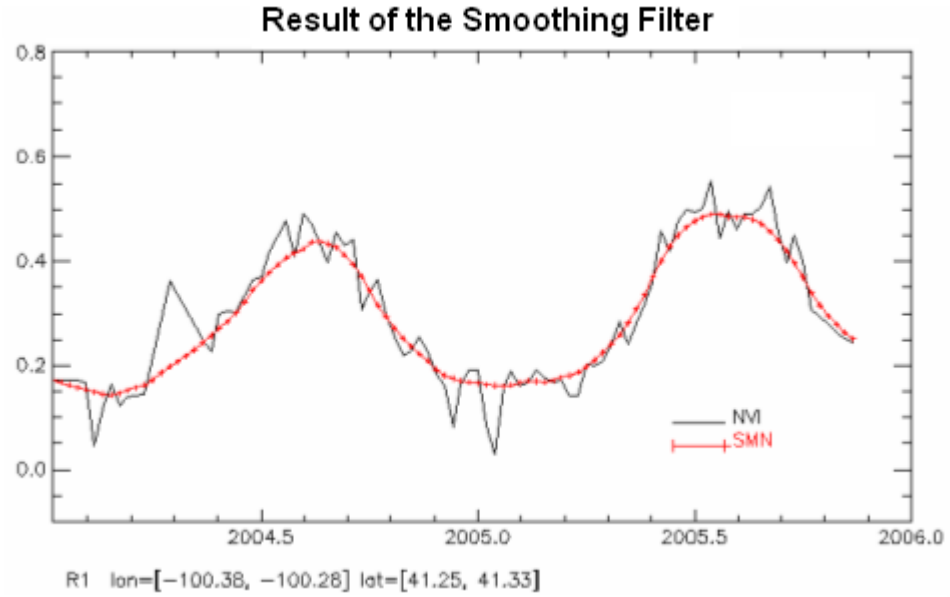
- Klasyczna regresja
- Wygładzanie („Smoothing”)
- Modele ARIMA
- Analiza spektrum

Regresja

- Podejście najbardziej banalne
- Wyjściem jest funkcja liniowa
- Estymuje trend

Wygładzanie (Smoothing)

- Uśrednianie wyników w celu zminimalizowania nagłych skoków danego szeregu
- Skupia się na jednym atrybucie
- Gubi informację dotyczącą dynamiki szeregu



Modele ARIMA

- Dopuszcza analizę wpływu danych historycznych na dane obecne
- Przeznaczona do stosunkowo krótkich serii
- Estymacja kolejnej wartości na podstawie *p* wartości poprzednich

$$x_t = 0.7\, x_{t-1} - 1.4\, x_{t-2} + \omega_t$$

- Prognozowanie wartości

Analiza spektrum

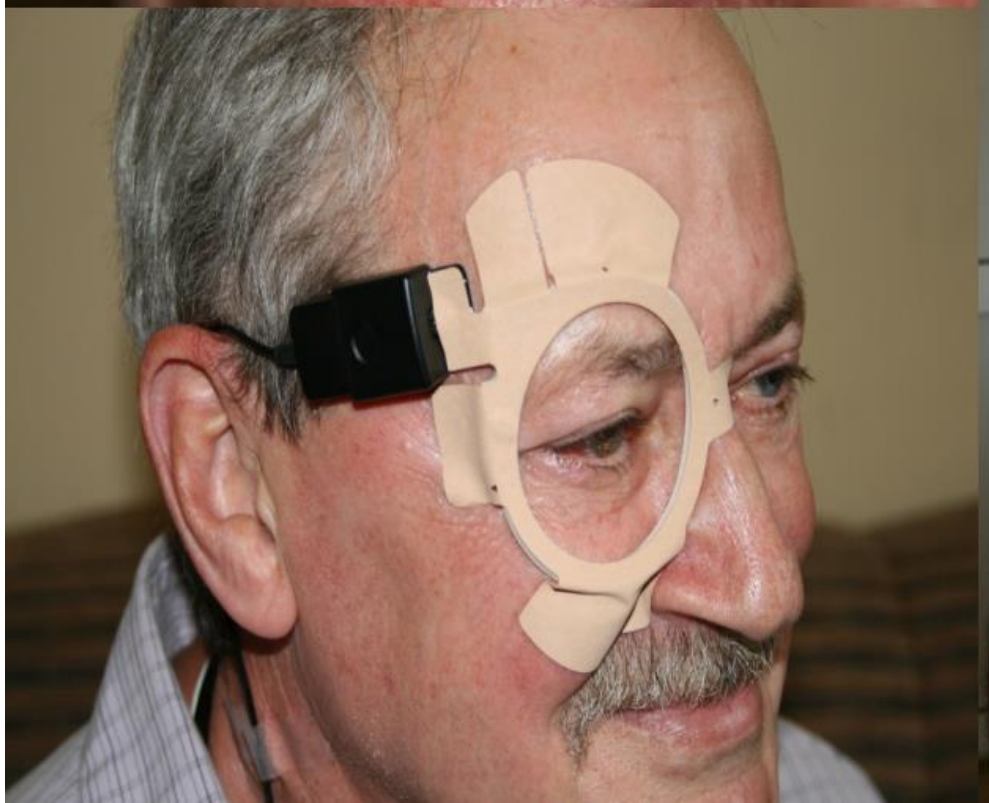
- Uwzględnia powtarzalność, okresowość
- Wiele rozszerzeń dopasowujących metodę do badanego szeregu
  - Transformaty Fouriera
  - Gęstość spektrum
  - Filtry liniowe
- Skomplikowana matematycznie

$$x_t = U_1 \cos(2\pi\omega_0 t) + U_2 \sin(2\pi\omega_0 t)$$

Problem wyjściowy

- Stworzenie klasyfikatora potrafiącego zdecydować, czy badany pacjent jest chory na jaskrę, czy nie.
- Jakie posiadamy dane?
  - Diagnoza: pacjent chory; pacjent zdrowy
  - Time: Czas wykonania badania
  - TFADJ: Dane z urządzenia

  Dane pobierane są z urządzenia co 15 min przez całą dobę



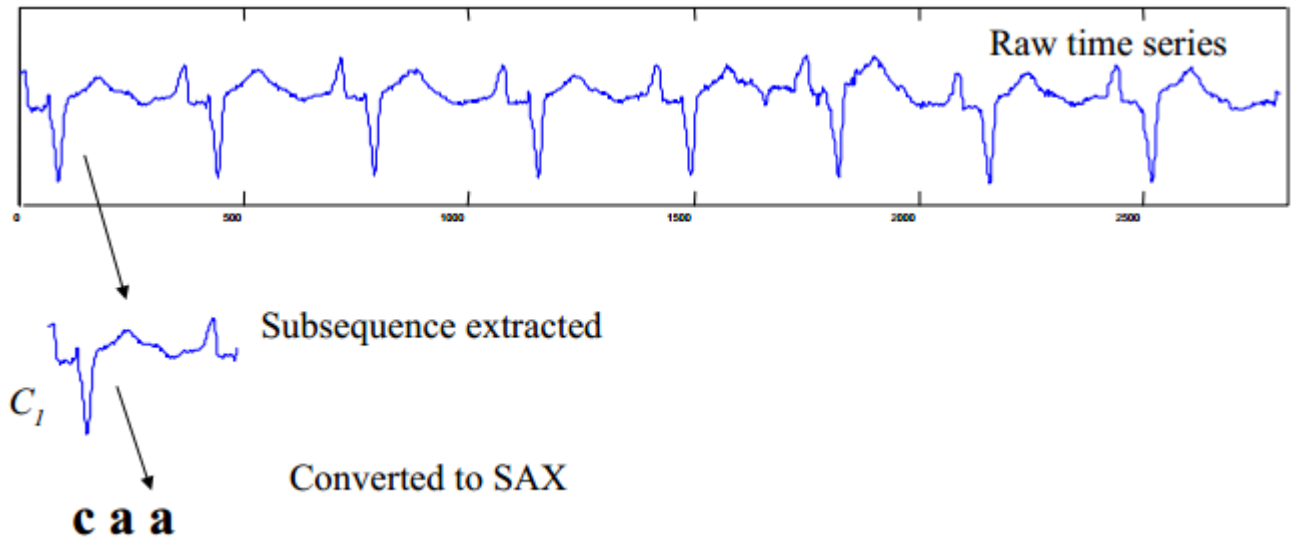
Podsumowanie omówionych metod

- Skupiają się raczej na predykcji niż na decyzji (klasyfikacji) – której to predykcji nie potrzebujemy
- Nie zawsze są wystarczająco dokładne (regresja)
- Ewentualna klasyfikacja byłaby skupiona na uśrednianiu, co ponownie wprowadza niedokładności
- Celem jest wykorzystanie wiedzy, który pacjent jest chory, a który zdrowy

Wniosek: Potrzebujemy czegoś innego

SAX - Symbolic Aggregate approXimation

- Eamonn Keogh oraz Jessica Lin (2002 r.)
- Zamiana szeregu czasowego w ciąg symboli
  - Redukcja wymiarowości
  - Standaryzacja długości
  - Dyskretyzacja



Mówię jak jest

- Rozszerzenie SAX do iSAX (indexed SAX) – Uniwersytet Kalifornijski
  - Przetwarzanie masywnych danych w sposób dokładny i wydajny
  - Metoda uznana przez statystyków – wspierana przez SIGKDD (Special Interest Group on Knowledge Discovery and Data Mining, stowarzyszenie ACM)

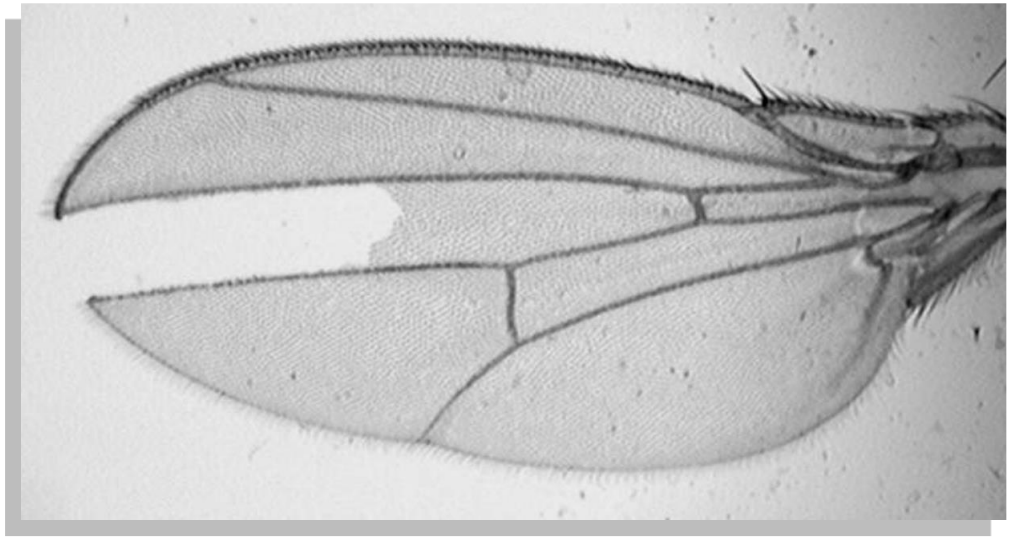
  „…we can find similarity searches using edit distance over 10,000 time series in 50 milliseconds.”

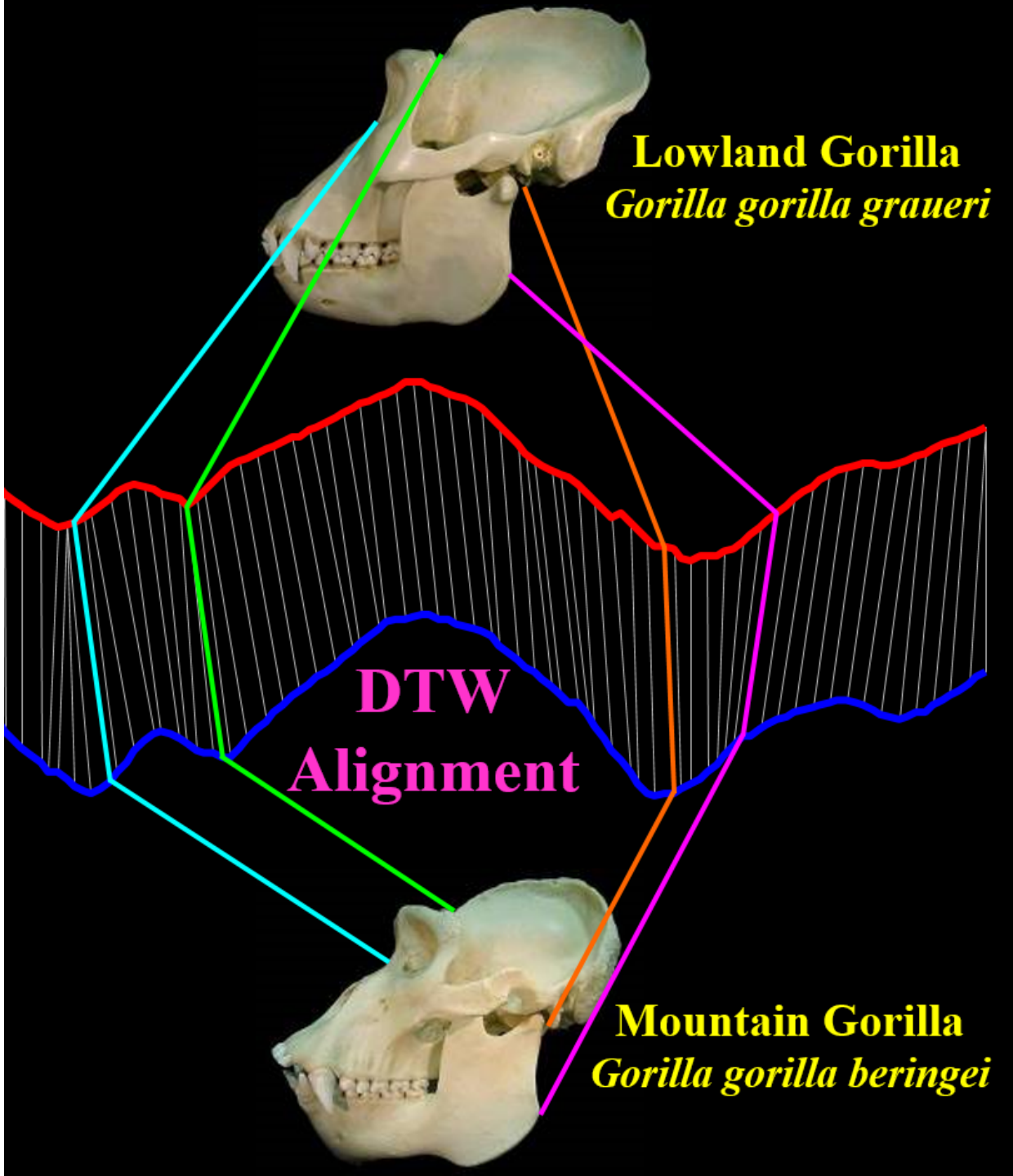
  Ray Cromwell, Timepedia.org

  *„In our current research the (**SAX**) symbolic representation of Lin and Keogh wins out even over well-known approximations...”*

  Data Mining Applications in the Automotive Industry. 2010 Kruse, Steinbrecher and Moewes

Badanie obrazów





Jak to wykorzystać ?

Dane szeregi czasowe

SAX

Zwykłe dane
(para: ciąg symboli; diagnoza)

Klasyfikator

Dziękuję za uwagę

- Bibliografia:
  - Shieh J., Keogh E., iSAX: Indexing and Mining Terabyte Sized Time Series, SIGKDD 2008
  - Shumway R., Stoffer D., Time Series Analysis and Its Applications, 2011